XAI4EXTREMES: AN INTERPRETABLE MACHINE LEARNING FRAMEWORK FOR UNDERSTANDING EXTREME-WEATHER PRECURSORS UNDER CLIMATE CHANGE

Jiawen Wei
National University of Singapore

Aniruddha Bora & Vivek Oommen
Brown University

Chenyu Dong
National University of Singapore

Juntao Yang & Jeff Adie
NVIDIA AI Technology Centre

Chen Chen
Centre for Climate Research Singapore

Simon See
NVIDIA AI Technology Centre

George Karniadakis
Brown University
george_karniadakis@brown.edu

Gianmarco Mengaldo
National University of Singapore
mpegim@nus.edu.sg

ABSTRACT

Extreme weather events are increasing in frequency and intensity due to climate change. This, in turn, is exacting a significant toll in communities worldwide. While prediction skills are increasing with advances in numerical weather prediction and artificial intelligence tools, extreme weather still present challenges. More specifically, identifying the precursors of such extreme weather events and how these precursors may evolve under climate change remain unclear. In this paper, we propose to use post-hoc interpretability methods to construct relevance weather maps that show the key extreme-weather precursors identified by deep learning models. We then compare this machine view with existing domain knowledge to understand whether deep learning models identified patterns in data that may enrich our understanding of extreme-weather precursors. We finally bin these relevant maps into different multi-year time periods to understand the role that climate change is having on these precursors. The experiments are carried out on Indochina heatwaves, but the methodology can be readily extended to other extreme weather events worldwide.

1 INTRODUCTION

Climate change is playing a pivotal role in exacerbating extreme weather across the globe, with extreme weather events becoming more frequent and severe (Masson-Delmotte et al., 2021). These events, in turn, are exerting heavy socioeconomic and environmental tolls on communities and fragile ecosystems worldwide. For instance, the tropical Indo-Pacific is witnessing an increased frequency of heatwaves and extreme precipitation due to critical changes in synoptic weather patterns in the region (Dong et al., 2024). Similarly, heatwaves during the summer and storms during the winter are becoming more frequent in Europe due to atmospheric circulation changes (Faranda et al., 2023). Other regions are also experiencing an increased frequency of heatwaves and other extremes – see for instance (Perkins-Kirkpatrick & Lewis, 2020; Donat et al., 2016).

Extreme weather is commonly defined as weather that is significantly different from the typical conditions for a particular region and time of the year. Heatwaves, for example, are due to prolonged periods of excessively hot weather relative to the expected conditions in a given area and time, that can lead to significant impacts on the affected community, such as health, infrastructure, and agricultural issues. In order to mitigate the impact of weather extremes, authorities typically issue

warnings ahead of time to alert the population that may be at risk. These warnings are largely based on weather forecasts provided by local and and global operational weather services. However, the prediction skills for certain types of extremes and regions is still relatively poor, for both short-range (a few days ahead) and mid-range (two weeks ahead) forecasts. One example is related to the forecast of heatwaves in large portions of the tropics, including the Maritime continent, and central Africa, as well as in the Carribean and Central America and the western US (De Perez et al., 2018).

Indeed, predictability drivers for heatwaves vary across different regions, and they are due to the confluence of several physical mechanisms, including diabatic heating from radiation and surface heat fluxes, adiabatic warming from air subsidence, and horizontal movement of hot air masses (De Perez et al., 2018). At mid-latitudes, heatwaves are frequently associated with persistent atmospheric blocking events, which promote subsidence and clear-sky conditions, thereby enhancing surface warming (Kautz et al., 2022). In the tropics, typical drivers are instead dominated by high solar radiation and reduced cloud cover that amplify surface heating, as well as by suppressed convection from large-scale subsidence, and warm sea surface temperature anomalies during e.g., El Niño events (Cai et al., 2014). In addition, land conditions and vegetation, such as land moisture levels and presence of forest vs grassland are important drivers of heatwaves Domeisen et al. (2023). These drivers help increasing the confidence in the forecast of heatwaves, providing an important time window to issue early warnings to the affected population. These drivers, also referred to as precursors, are typically the result of human-expert knowledge, or briefly the "*human view*". In this work, we take a different perspective, and look at these precursors through the lenses of interpretable machine learning (ML), thereby providing a possibly complementary "*machine view*". The latter is obtained by identifying what data the machine deemed important to the onset of heatwaves, and it is used to understand (by working with human domain experts) whether it may be helpful in enriching our understanding of precursors – see also Mengaldo (2024) for the use of explainable artificial intelligence (XAI) for scientific knowledge discovery. Without losing generality in the methodology proposed, we focus on tropical heatwaves in the Indochina peninsula, and attempt to answer two questions via interpretable ML: (i) What are the key precursors of these events? (ii) Is climate change influencing these precursors?

2 METHODOLOGY

To outline our approach, we focus on heatwaves in the Indochina peninsula (the latter depicted in Figure 3). These can be divided into heatwaves in the dry and in the wet seasons, whereby the precursors and onset mechanisms differ (Luo & Lau, 2018). We focus on dry-season (FMAM) heatwaves without lacking generality on the methodology proposed here.

The key idea is to look at these extreme weather events, namely dry-season heatwaves, using interpretable machine learning; more specifically post-hoc interpretability methods applied to a binary time series classification deep learning (DL) framework. This approach allows producing relevance maps, that highlight what input data the DL framework deemed important for the prediction it made. The binary DL time series classification framework is setup as follows. As input data, we consider the spatial (i.e., geographical) maps of 23 variables for the 7 days prior of a heatwave striking the Indochina peninsula. The 23 input variables characterize the large majority of dry-season heatwave precursors, and the 7 days time window provides a relevant time frame to capture the underlying pathways leading to these extremes. We then assume that the DL framework is able to identify patterns in the data that are causal to heatwaves; in other words, we assume that it could capture systematically the precursors to heatwaves. Indeed, we consider only true positive samples, such that the data deemed important by the DL framework, also referred to as "*machine view*", is only associated to correctly classified heatwaves. The binary labels for the classification task are (1) heatwave and (0) non-heatwave, where the heatwaves are identified as outlined in Appendix A.

The final heatwave binary classification dataset consists of 720 samples with an approximate ratio of (1) heatwave vs (0) non-heatwave being 1:5. We split the dataset into training, validation, and testing sets with a ratio of [0.6:0.2:0.2], and then train the Transformer model for heatwave classification. We apply four different post-hoc interpretability methods, namely Integrated Gradients (Sundararajan et al., 2017), DeepLIFT (Shrikumar et al., 2017), DeepSHAP (Lundberg, 2017), and GradSHAP (Lundberg, 2017), to the trained Transformer model. To guarantee that we obtain the most accurate and robust relevance maps, we adopt the interpretability evaluation frameworks in

Turbé et al. (2023) and Wei et al. (2024). Integrated Gradients performs the best among four post-hoc methods according to the evaluation results; thereby we use the relevance maps it generates for analysis. The overall approach, that we name XAI4Extremes, is depicted in Figure 1: we propose a new dataset for weather extremes – heatwaves in this particular case (panel a, in gray), that is used by a predictive DL framework (panel b, in blue), to which we apply post-hoc interpretability and its evaluation (panel c, in red). The relevance maps produced by the post-hoc interpretability method, what we also refer to as "*machine view*" (panel d, in red), are then compared against human expert knowledge, what we also refer to as "*human view*" (panel e, in green). This comparison may lead to knowledge discovery in terms of heatwave precursors and role of climate change in heatwave precursors. This may be the case when the machine view enriches human expert knowledge, by providing a scientifically plausible use of data that was unknown to human domain experts, but that domain experts can explain. Indeed, it is responsibility of human domain experts to respond to the question **why** the interpretable machine learning framework deemed important a specific set of input data. The relevance maps can also be used to generate adversarial samples to augment the dataset and shape model behavior, thereby improving the performance of the predictive DL framework. We remark that the approach outlined in this section can readily be applied to other types of weather extremes in different regions worldwide.

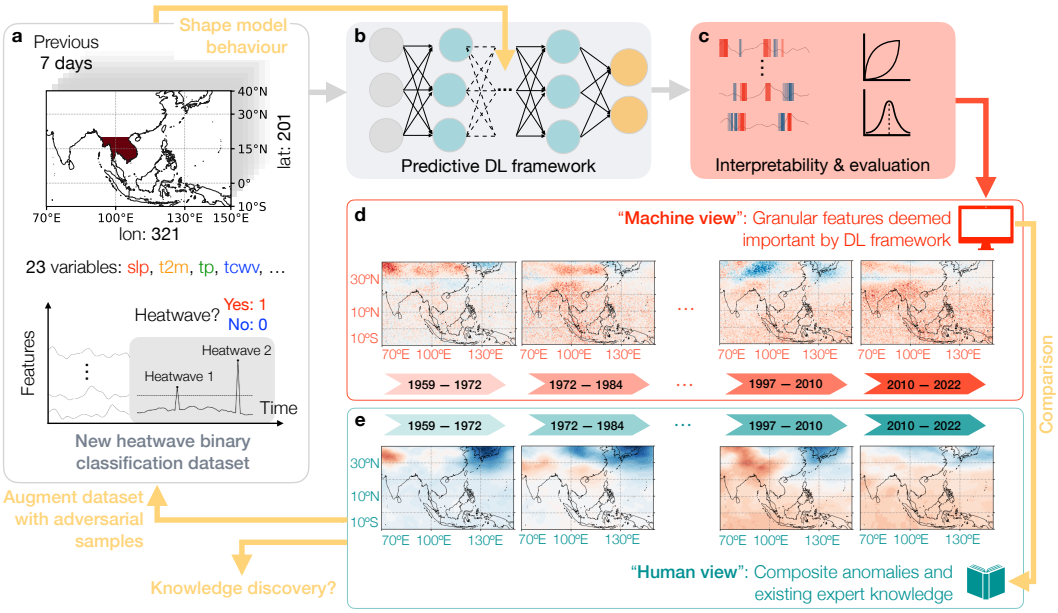

Figure 1: The XAI4Extremes framework proposed, composed of a novel extreme weather dataset (a), a DL predictive model (b), an interpretability block along with its evaluation (c), that produces relevance maps, or what we called the "*machine view*" (d). The latter (d) is then compared with existing human expert knowledge (e) for knowledge discovery or for augmenting the dataset with e.g., adversarial samples that can shape and improve model behavior.

3 Results

Figure 2 shows the temperature field at 200 hPa, that is the temperature between approximately 11 and 12 km altitude (i.e., the temperature in the upper troposphere), for two different regions, region 1 and 2. Region 1 comprises the Indian Ocean, and India, while region 2 comprises the Maritime continent and part of the Pacific Ocean. In Figure 2, panel a, we show the mean trend of relevance for region 1 (top row), region 2 (middle row), and region 1 and 2 combined (bottom row). It is possible to see how the temperature in the upper troposphere is deemed more important by the machine for heatwaves in Indochina in more recent decades for both regions, with a clear upward trend. If we compare the interpretability results (i.e., the relevance maps or machine view) with something more

understandable by humans, i.e., composite anomalies, we note that there is indeed a warming of the upper troposphere that is associated to heatwaves in Indochina (Figure 2, panel b). This indicates that the temperature at 200 hPa is becoming a key precursor of Indochina heatwaves, especially in recent decades (in agreement with composite anomalies – see Figure 10 in Appendix B.3), aspect that may indicate the fingerprint of climate change. If we further compare the contribution of the

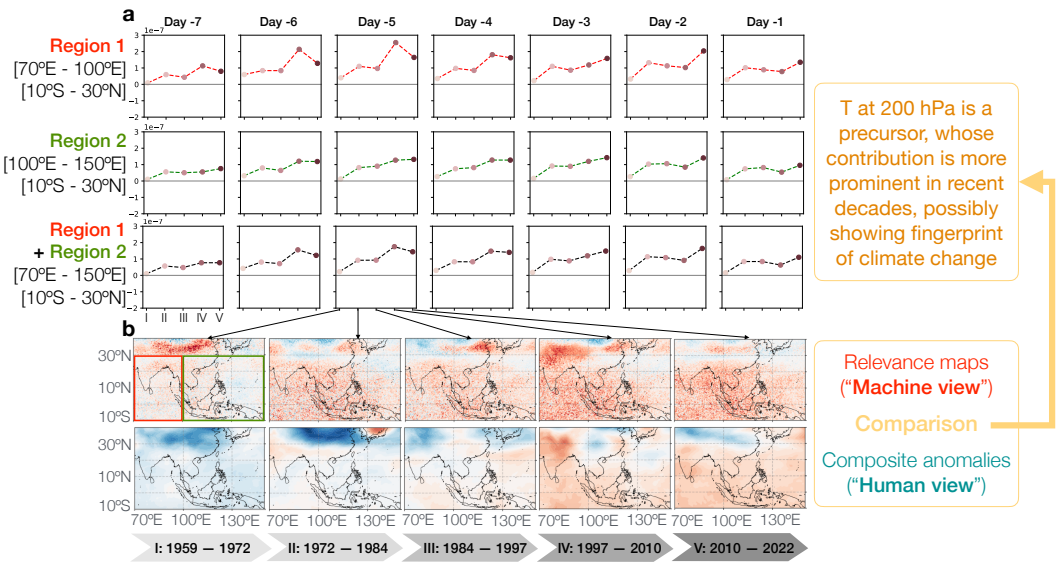

Figure 2: Mean relevance of temperature at 200 hPa for the 5 historical time periods considered and on the 7 days prior to heatwaves in Indochina, for region 1 (a, top), region 2 (a, middle), region 1+2 (a, bottom), along with the corresponding relevance maps – i.e., "*machine view*" – associated to 7 days prior to heatwave, and composite anomalies – i.e., "*human view*" – (b).

other 22 variables, we observe how the temperature in the upper troposphere is one of the most important variables for the DL prediction – see Appendix B.2, i.e., Figures 4 to 9. In addition, if we look at the contribution of two common variables related to climate warming, namely 2-meter temperature and max temperature (Appendix B.3, Figures 11 and 12, respectively), we observe how these have a strong trend in terms of composite anomalies, but not in terms of relevance maps. This result points to a human-understandable explanation where higher 200 hPa temperature can suppress convection and increase subsidence, thereby leading reduced cloud cover that amplifies surface heating, potentially leading to heatwaves. Onset mechanism that may have been exacerbated by climate change, with a fingerprint on the 200 hPa temperature. In Appendix B.1, Table 1, we also show the predictive performance of the DL model for the classification task, on the 5 historical time periods considered. We observe that the performance can be significantly improved, aspect that is currently ongoing, with some preliminary results provided in Table 2.

4 Conclusions

The preliminary results presented in this work aim to introduce the overarching explainable AI framework we propose, namely XAI4Extremes. The framework has the key objective of better understanding weather extremes and their evolution under climate change. To achieve this task, we create a novel binary classification dataset for heatwaves in Indochina (noting that the same dataset type can be created for other weather extremes in other regions worldwide). We then propose to couple a predictive DL framework with interpretability methods, in order to understand ***what*** data the machine deemed important for its predictive performance of true positive samples (i.e., correctly identified heatwaves), something we refer to as "*machine view*". We finally propose to compare this machine view to existing human expert knowledge (what we call "*human view*"), to respond the question ***why*** the machine used those data. The latter aspect may lead to knowledge discovery, or it can be used to shape model behavior by e.g., generating ad-hoc adversarial samples based on the

machine view. We note that there are still several, yet stimulating, open challenges to be overcome. For instance, how to further improve the robustness of post-hoc interpretability methods, how to develop effective ante-hoc (also referred to as self-explainable) ML approaches (e.g., Turbé et al., 2024) for spatio-temporal data, how to use relevance maps for spatio-temporal data, and isolate features that are relevant, how to guarantee that the patterns identified by the relevance maps are causal to the task, among others. We believe that these limitations are open opportunities for the AI and broader scientific research communities that can be tackled over the next few years.

ACKNOWLEDGMENTS

J.W. and G.M. acknowledge support from MOE Tier 2 grant T2EP50221-0017, and from MOE Tier 1 grant 22-4900-A0001-0.

REFERENCES

Wenju Cai, Simon Borlace, Matthieu Lengaigne, Peter Van Rensch, Mat Collins, Gabriel Vecchi, Axel Timmermann, Agus Santoso, Michael J McPhaden, Lixin Wu, et al. Increasing frequency of extreme el niño events due to greenhouse warming. *Nature climate change*, 4(2):111–116, 2014.

Erin Coughlan De Perez, Maarten Van Aalst, Konstantinos Bischiniotis, Simon Mason, Hannah Nissan, Florian Pappenberger, Elisabeth Stephens, Ervin Zsoter, and Bart Van Den Hurk. Global predictability of temperature extremes. *Environmental Research Letters*, 13(5):054017, 2018.

Daniela IV Domeisen, Elfatih AB Eltahir, Erich M Fischer, Reto Knutti, Sarah E Perkins-Kirkpatrick, Christoph Schär, Sonia I Seneviratne, Antje Weisheimer, and Heini Wernli. Prediction and projection of heatwaves. *Nature Reviews Earth & Environment*, 4(1):36–50, 2023.

Markus G Donat, Andrew L Lowry, Lisa V Alexander, Paul A O'Gorman, and Nicola Maher. More extreme precipitation in the world's dry and wet regions. *Nature Climate Change*, 6(5):508–513, 2016.

Chenyu Dong, Robin Noyelle, Gabriele Messori, Adriano Gualandi, Lucas Fery, Pascal Yiou, Mathieu Vrac, Fabio D'andrea, Suzana J Camargo, Erika Coppola, et al. Indo-pacific regional extremes aggravated by changes in tropical weather patterns. *Nature Geoscience*, pp. 1–8, 2024.

Davide Faranda, Gabriele Messori, Aglae Jezequel, Mathieu Vrac, and Pascal Yiou. Atmospheric circulation compounds anthropogenic warming and impacts of climate extremes in europe. *Proceedings of the National Academy of Sciences*, 120(13):e2214525120, 2023.

Lisa-Ann Kautz, Olivia Martius, Stephan Pfahl, Joaquim G. Pinto, Alexandre M. Ramos, Pedro M. Sousa, and Tim Woollings. Atmospheric blocking and weather extremes over the euro-atlantic sector–a review. *Weather and Climate Dynamics*, 2022.

Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Ming Luo and Ngar-Cheung Lau. Synoptic characteristics, atmospheric controls, and long-term changes of heat waves over the indochina peninsula. *Climate Dynamics*, 51:2707–2723, 2018.

Valérie Masson-Delmotte, Panmao Zhai, Anna Pirani, Sarah L Connors, Clotilde Péan, Sophie Berger, Nada Caud, Y Chen, L Goldfarb, MI Gomis, et al. Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, 2(1):2391, 2021.

Gianmarco Mengaldo. Explain the black box for the sake of science: Revisiting the scientific method in the era of generative artificial intelligence. *arXiv preprint arXiv:2406.10557*, 2024.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

SE Perkins, LV Alexander, and JR Nairn. Increasing frequency, intensity and duration of observed global heatwaves and warm spells. *Geophysical Research Letters*, 39(20), 2012.

SE Perkins-Kirkpatrick and SC Lewis. Increasing trends in regional heatwaves. *Nature communications*, 11(1):3357, 2020.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMlR, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Hugues Turbé, Mina Bjelogrlic, Christian Lovis, and Gianmarco Mengaldo. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence*, 5(3):250–260, 2023.

Hugues Turbé, Mina Bjelogrlic, Gianmarco Mengaldo, and Christian Lovis. Protos-vit: Visual foundation models for sparse self-explainable classifications. *arXiv preprint arXiv:2406.10025*, 2024.

Jiawen Wei, Hugues Turbé, and Gianmarco Mengaldo. Revisiting the robustness of post-hoc interpretability methods. *arXiv preprint arXiv:2407.19683*, 2024.

A IDENTIFICATION OF HEATWAVES

Identifying heatwaves remains a significant challenge. Currently, there are numerous definitions of heatwaves in the research community, yet there is no consensus on a standard definition. This complexity arises from the varied spatial coverage and duration of heatwaves. In our study, we adopted a relatively simple two-stage definition that combines index-based and event-based approaches, which have been widely used in other research.

We first define heatwaves on each individual grid point in the daily ERA5 reanalysis data from 1959 to 2022 using the heatwave index TX90pct (Perkins et al., 2012). The threshold for one day at one grid point is the calendar day 90th percentile of the daily maximum temperature, based on a centered 15-day window. A heatwave is defined as three or more consecutive days exceeding this threshold, and all days belonging to this heatwave are considered as heatwave days for that grid point. We note that we removed a grid point by grid point linear trend from the the data. This is because we want to maintain a relatively uniform distribution of heatwaves in the studied period.

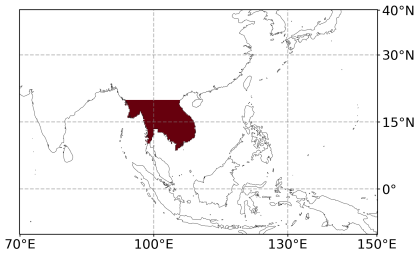

Figure 3: Indochina region used to define heatwaves (dark red).

Based on this grid point by grid point definition of local heatwaves, we further define heatwave events in Indochina using the regional mask illustrated in figure 3. For each region, one heatwave event is defined when a minimum number of grid points are identified as heatwaves. Specifically, this threshold is set at the 90th percentile of the number of grid points classified as heatwaves during the season of interest. We define the first day that exceeds the predefined threshold as the heatwave onset day. To avoid overlapping events, we stipulate that no day within the seven days preceding any onset day should exceed this threshold. For the onset days of non-extreme events, we randomly select days when the number of grid points falls below the predefined threshold within the same

season, following specific criteria: We ensure that there are no heatwave onset days or other non-extreme events within a 7-day window before and after these selected days. We provide the dataset with a ratio of non-extreme events to extreme events set at 5:1. This is the maximum ratio achievable while following the selection strategy outlined above.

B ADDITIONAL RESULTS

B.1 PREDICTIVE PERFORMANCE OF DL FRAMEWORK

In Table 1, we show the predictive performance of the Transformer model used as a reference for this study. We observe that the performance can be significantly improved, something that is currently ongoing. To this end, in Table 2, we also present the results (for the testing set only) for Convolutional Encoder with Attention blocks (Conv+Attn) and FourCastNet (Pathak et al., 2022) models that are being rolled out. The Conv+Attn model used in this study consists of three sets of convolution and maxpool operations (that down-samples the latent representation by a factor of 2) resulting in a latent representation at 1/8th of the input resolution. The first convolution block contains a channel-wise attention layer and the last convolution block is followed by a spatial attention block. The output of the last convolutional block is projected to a scalar using a fully connected layer with sigmoid activation, that predicts if whether or not the heat wave is going to occur. The FourCastNet used here is modified from the original code (Pathak et al., 2022) by adding a extra block specifically for the binary classification of the heatwaves.

Table 1: Predictive performance of trained Transformer model

	Sensitivity (TPR)	Specificity (TNR)	Precision (PPV)	Miss Rate (FNR)	Accuracy
1959 — 1972	52.38%	100.00%	100.00%	47.62%	93.06%
1972 — 1984	35.48%	100.00%	100.00%	64.52%	86.11%
1984 — 1997	57.69%	99.15%	93.75%	42.31%	91.67%
1997 — 2010	50.00%	100.00%	100.00%	50.00%	93.75%
2010 — 2022	66.67%	99.17%	94.12%	33.33%	93.75%

Table 2: Predictive performance of other models being rolled out – results shown are for the testing period only, that is between 26 Feb 2010 to 18 Dec 2022.

	Sensitivity (TPR)	Specificity (TNR)	Precision (PPV)	Miss Rate (FNR)	Accuracy
Conv+Attn	70.83%	96.67%	80.95%	29.16%	92.36%
FourCastNet	70.83%	97.50%	85.00%	29.16%	93.75%

B.2 MEAN RELEVANCE FOR ALL VARIABLES

In this section, we present the mean relevance for all the 23 variables considered, noting how the temperature at 200 hPa is among the variables that is deemed most important by the DL framework for predicting heatwaves in Indochina. In particular, Figure 4 shows the mean relevance, including positive and negative, for the 23 variables considered in this work, on the 7 days prior to heatwaves, and across the 5 time periods taken into account, for region 1. Figure 5 is equal to Figure 4, except that it only considers positive relevance.

Figures 6, 7, and Figures 8, 9 show the same quantities as Figures 4, 5 for region 2, and for region 1 and region 2 together, respectively.

Figure 4: Mean relevance, **positive and negative**, on **region 1**, for all 23 variables considered, on the 7 days prior to heatwaves in Indochina and for the 5 historical time periods considered.

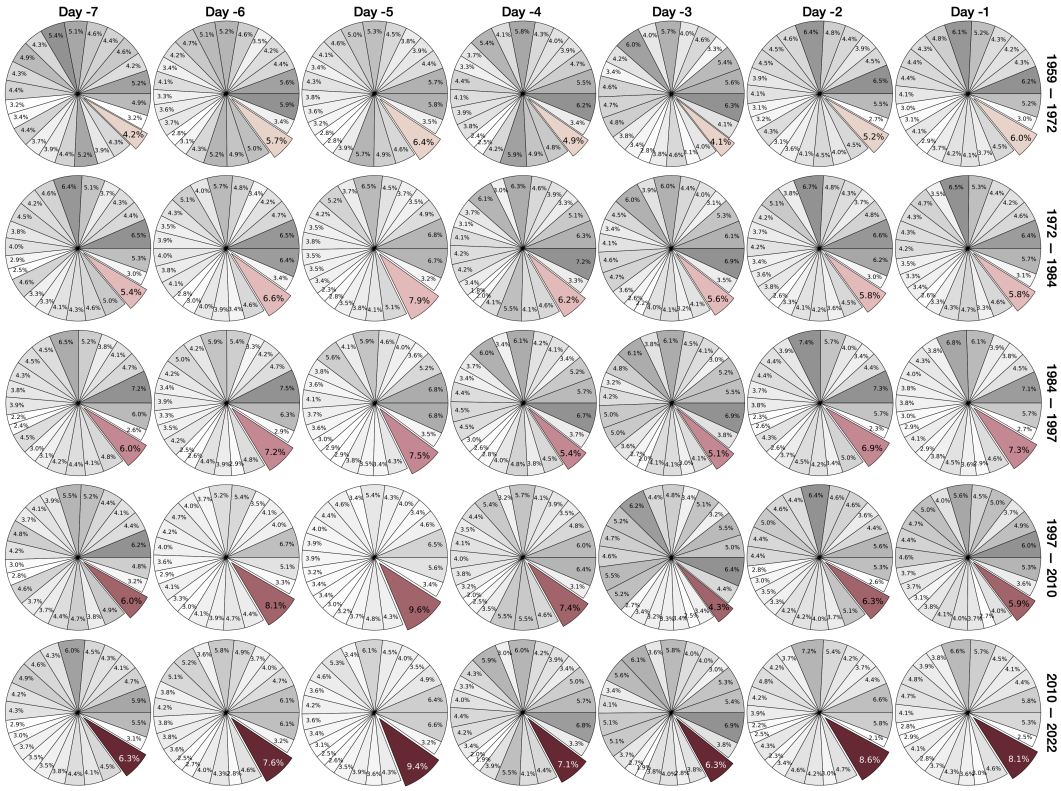

Figure 5: Mean relevance, **only positive** (negative is set to zero) on **region 1**, for all 23 variables considered, on the 7 days prior to heatwaves in Indochina and for the 5 historical time periods considered. Variable t_200hPa (temperature at 200hPa) is highlighted with red colors across 5 historical time periods, while the other variables are displayed in gray.

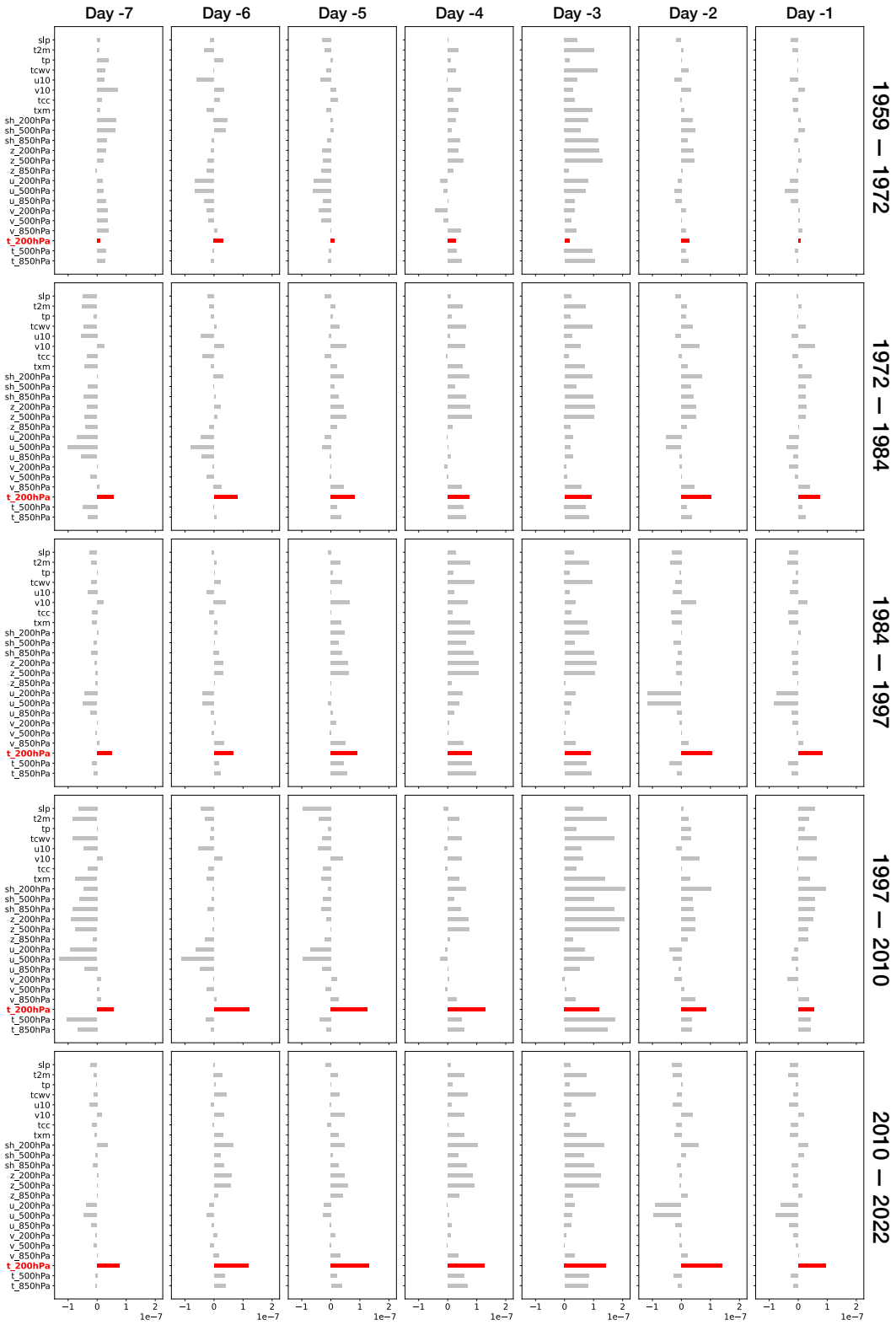

Figure 6: Mean relevance, **positive and negative**, on **region 2**, for all 23 variables considered, on the 7 days prior to heatwaves in Indochina and for the 5 historical time periods considered.

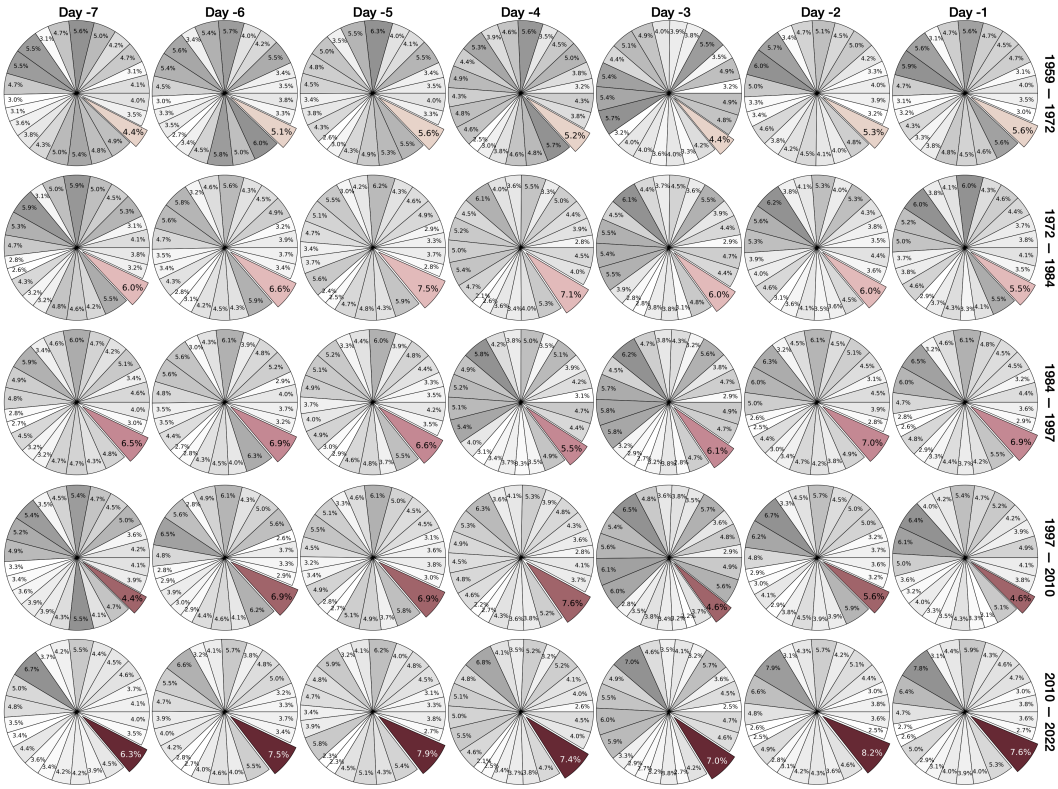

Figure 7: Mean relevance, **only positive** (negative is set to zero) on **region 2**, for all 23 variables considered, on the 7 days prior to heatwaves in Indochina and for the 5 historical time periods considered. Variable t_200hPa (temperature at 200hPa) is highlighted with red colors across 5 historical time periods, while the other variables are displayed in gray.

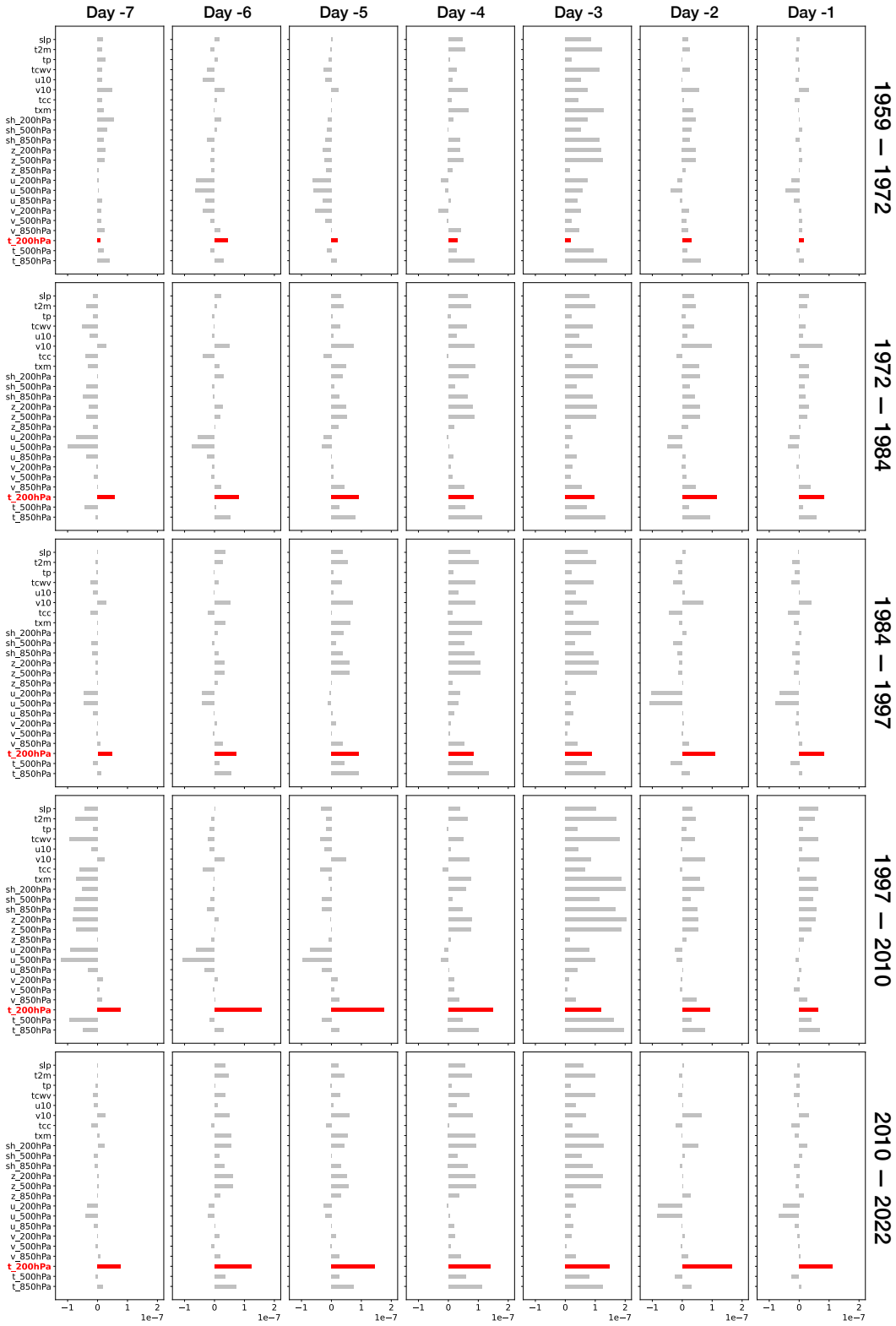

Figure 8: Mean relevance, **positive and negative**, on **region 1 + region 2**, for all 23 variables considered, on the 7 days prior to heatwaves in Indochina and for the 5 historical time periods considered.

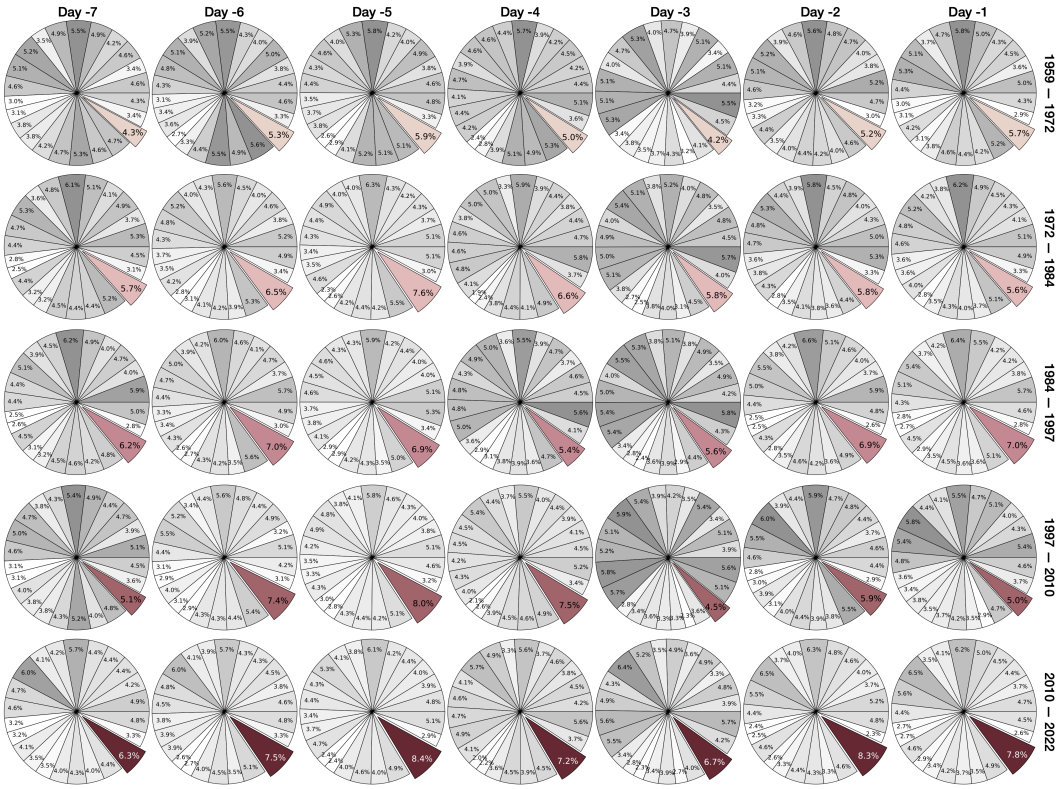

Figure 9: Mean relevance, **only positive** (negative is set to zero) on **region 1 + region 2**, for all 23 variables considered, on the 7 days prior to heatwaves in Indochina and for the 5 historical time periods considered. variable t_200hPa (temperature at 200hPa) is highlighted with red colors across 5 historical time periods, while the other variables are displayed in gray.

B.3 RELEVANCE VS COMPOSITE ANOMALIES

In this section, we present the comparison between the relevance maps and the composite anomalies for thre three regions considered, and for three key variables related to climate change, namely: 200 hPa temperature (Figure 10), 2-meter temperature (Figure 11), and for maximum temperature (Figure 12). The results show an upward trend in terms of composite anomalies for all variables and nearly all regions. However, relevance exhibits an uptrend only for 200 hPa temperature.

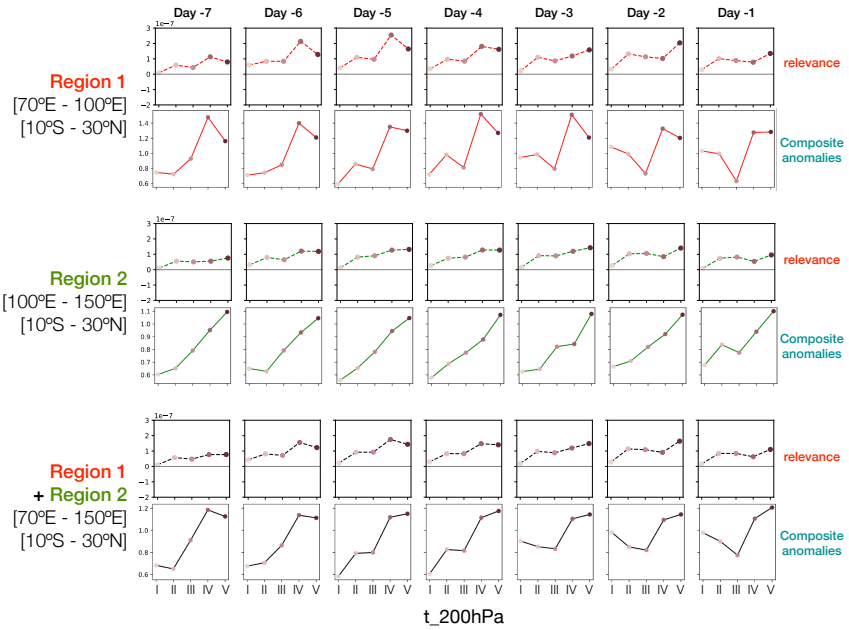

Figure 10: Mean trend of relevance maps vs composite anomalies for variable t_200hPa (temperature at 200hPa) on region 1 (top row), region 2 (middle row), and region 1 and 2 combined (bottom row).

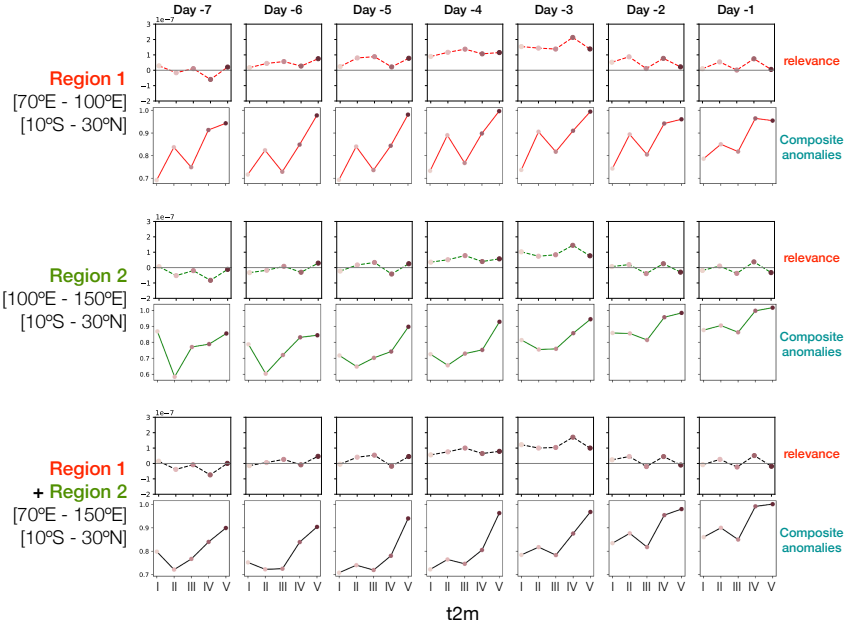

Figure 11: Mean trend of relevance maps vs composite anomalies for variable t2m (2-meter temperature) on region 1 (top row), region 2 (middle row), and region 1 and 2 combined (bottom row).

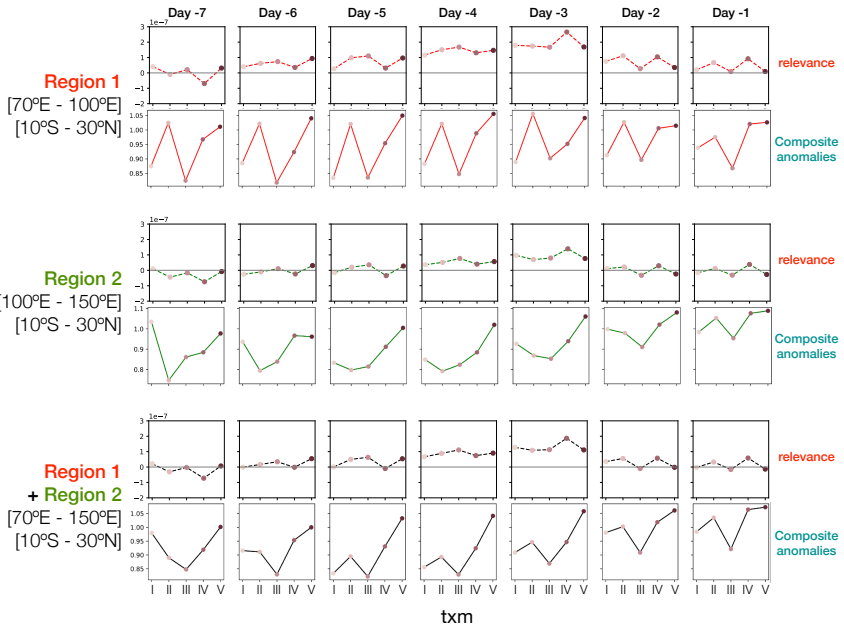

Figure 12: Mean trend of relevance maps vs composite anomalies for variable txm (maximum temperature) on region 1 (top row), region 2 (middle row), and region 1 and 2 combined (bottom row).